DOCUMENT RESUME

ED 209 653 CS 006 367

AUTHOR Schell, Leo M.
TITLE How Accurate Are Oral Reading Tests?
PUB DATE Oct 81
NOTE 14p.; Paper presented at the Annual Meeting of the Plains Regional Conference of the International Reading Association (9th, Des Moines, IA, October 22-24, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; Informal Reading Inventories; *Oral Reading; *Reading Tests; Scores; Scoring; Standardized Tests; *Testing Problems; *Test Reliability; *Test Validity

ABSTRACT

Errors in oral reading tests result from inaccuracies that tend to creep in because children are not totally consistent while taking a test and from inaccuracies caused when the examiner does not catch a word recognition error, giving credit for an answer that is more wrong than right or vice versa. Every test contains a standard error of measurement (SEM), indicating how much an individual's score is likely to fluctuate on repeated testings. Some standardized oral reading tests also report this, but SEMs are almost totally ignored by authors of informal or standardized oral reading tests. An examination of recently published informal reading inventories revealed that three of five did not mention that the obtained score was not totally accurate and dependable. In the other two, only a sentence or two at the end of the manual hinted that the results might need to be verified. The same also is true of two recently revised standardized oral reading tests. Studies have shown how subjective the scoring of oral reading tests is and the amount of examiner error these scores contain. The score one receives definitely depends upon who is doing the scoring. Reading instructors should abandon the idea that a score on an oral reading test represents a point on a continuum and substitute for it the more reasonable notion that the score lies within a band of scores at least a grade level in range. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209653

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Leo M. Schell
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Leo M. Schell

HOW ACCURATE ARE ORAL READING TESTS?

The report from the remedial reading teacher said, "This child's Instructional Level, as obtained from an IRI, is 3.5" How accurate are the results of oral reading tests such as IRIs? How precisely do they measure children's reading achievement?

There are several sources of error in the results of all tests. One source is examinee error, that is, inaccuracies that tend to creep in because children aren't totally consistent while taking a test. Sometimes they'll make a lucky guess, other times they'll have a memory lapse or not understand the question. Sometimes such error produces higher scores than what they should really obtain while at other times it results in lower scores. This kind of error is statistically reported as test reliability and standard error of measurement.

Another source of error with oral reading tests is examiner error, that is, inaccuracies caused when the examiner doesn't catch a word recognition error, gives credit for an answer that is more wrong than right, etc. This kind of error has the possibility of occurring frequently in oral reading tests because they are subjectively scored.

CS006367

It is the contention of this article that neither of these kinds of error in oral reading tests--whether informal (IRIs) or standardized--has received the attention they deserve and that oral reading test scores are much less precise than we have typically treated them as being.

Standard Error of Measurement

Every test contains a certain amount of error, what Fry graphically refers to as "slop." He describes what this "slop" is.

> Statisticians call this slop the Standard Error of Measurement. It means how much an individual's score is likely to jump around on repeated testings. You see the problem is, "if a student scores high one day and lower the next day, on the same test, which score is correct?" Well, theoretically, if you retook the test many times and there was no practice effect, you could average the scores and get a "true score," which is the score you really want. But you can't give the test a lot of times in the real world; so you give it just once and estimate how far away you could be from the true score. This distance is called the Standard Error of Measurement (SEM) . . .

The signi icance of the SEM is that it tells how accurate the test score is. You take a child's score, for example, 25 raw score points, and you add and subtract the SEM, which let's say is 1.5 raw score points, and you get a band of probability from 23.5 to 26.5. This means that two times out of three, a child with a raw score of 25 would have a true score between 23.5 and 26.5. And it also means that in one out of three testings the true score would fall either above or below the band.

Some standardized oral reading tests report this SEM. The following SEMs and their probability bands were selected from two such tests, the Gray

Oral Reading Test (1963) and Diagnostic Reading Scales (1972), for illustrative purposes.

Table 1

Selected SEMs and Probability Bands
for the Gray Oral Reading Test

Grade Level	SEM (in Grade Equivalent units	Probability Band
2.0	.48	1.52 - 2.48
3.4	.54	2.86 - 3.94
4.8	.69	4.11 - 5.49
6.2	.76	5.44 - 6.98

Table 2

Selected SEMs and Probability Bands
for the Diagnostic Reading Scales

Grade Level	SEM (in Grade Equivalent units)	Probability Band
2.3	.3	2.0 - 2.6
4.5	.5	4.0 - 5.0
6.5	.6	5.9 - 7.1

Two aspects should be noted. One, the probability band covers a relatively wide span of reading levels, nearly a full grade level for second grade passages and up to 1 1/2 grade levels for sixth grade passages. The significance of this fact is that there is a considerable amount of error in oral reading tests--whether informal (IRIs) or standardized--and therefore their results should be interpreted as occurring within a range of achievement. Results should not be taken literally as precise points.

Two, SEMs--and their accompanying bands of probability--increase as grade (readability) level of the passages increases, e.g., the score of a child whose instructional level is measured as "6th grade" will be somewhat more inaccurate than that of a child whose instructional level is measured as "2nd grade."

However, SEMs are nearly totally ignored by authors of informal or standardized oral reading tests. An examination of the manuals of five recently published IRIs (Burns and Roe, 1980; Ekwall, 1979; Johns, 1981; Silvaroli, 1976; Woods and Moe, 1981) revealed that in three of them, not a single mention was made that the obtained score wasn't totally accurate and dependable. In the other two, only a sentence or two at the end of the manual hinted that error could creep in and that the results might need to be verified. The impression given was that the obtained score was the child's true score.

The same is also true of one of the two recently revised standardized oral reading tests, the Durrell Analysis of Reading Difficulty (1980). Even though the Durrell was standardized on 1224 students and norms were calculated, neither the term "standard error or measurement" nor the concept of test

inaccuracy is even mentioned in the manual. Spache's Diagnostic Reading Scales, both the 1972 and the 1981 editions, repeatedly use the word "estimate" as well providing such error statistics such as standard deviations, standard errors of measurement, and reliability coefficients. However, the manual never directly states the probability band for any specific passage level and, in general, leaves the impression that the test results are highly accurate when in fact Table 2 show that for instructional level passages at the 4th grade level and above the probability band was at least full grade level wide.

Interscorer Reliability

When administering an oral reading test, the examiner is required to make numerous instant decisions. Many of these decisions either do not have solid professional consensus (e.g., are repetitions counted as errors?) or must be highly subjective (e.g., was a word defined adequately to merit full credit?). This means that one examiner may not score the same test identically from one time to the next or that two examiners may not score the same test identically.

Some studies have investigated this problem. Ladd (1961) found that a group of teachers, after 30 hours of training with tape recorded tests or actual children or both, were still missing 33-37 percent of the errors made by children in oral reading tests. Page and Carlson (1975) had 17 well qualified reading specialists score and interpret a single 133-word tape-recorded passage of a child reading orally. Six of them estimated the material to be at the reader's independent level, five at the instructional level, and the remaining six at the frustration level. Bradley, Hanna, and

Lucas (1980) found that examiner errors in scoring the WISC-R could easily vary over a range of 6 to 8 IQ points and they concluded that "Clearly the score one receives depends in part upon who is doing the scoring." Obviously examiner error is widely prevalent, can be fairly large, and can have serious consequences.

A study was designed to examine scorer differences on the protocols of two separate reading test passages which were patterned very closely after the way these passages were read by two real children. One passage was chosen because it seemed straightforward and uncomplicated to score. The other was chosen because, in my classes, it had proven difficult to score, primarily because substitutions during oral reading affected the reader's responses to comprehension questions. It also contained several oral reading errors commonly made by remedial readers but whose scoring is not addressed in any manual or textbook known to me, e.g., one word miscalled twice in two separate sentences.

There were two groups of participants. Personal letters were sent to 50 professors listed in Graduate Programs and Faculty in Reading (Seifert, 1978). Twenty four of them, all from different institutions in 15 different states and Canada responded, of whom only three said they had given fewer than 50 oral reading tests. Eighteen--75 percent--said they had given more than 100 of them.

Seventy-three students enrolled in graduate reading courses at universities in Delaware, Pennsylvania, Tennessee, and Kansas also participated. All but 11 had nine or more hours in reading education and 45 percent of them had given more than 10 oral reading tests with 25 having given over 20 of them.

To maximize uniformity, all miscues were numbered in the protocol and participants were told which kinds of miscues constituted errors. The answers suggested in the manual were given along with the readers' responses. All necessary calculations were put into tables so that only circling a number was necessary rather than having to do any arithmetic.

The participants compared their scores with a standard IRI criteria table and determined whether the passage was at the reader's Independent, Instructional or Frustration level.

Table 3 reports the results for each passage for each of these two groups.

- -

INSERT TABLE 3 ABOUT HERE

- -

Conclusions

These results graphically illustrate how subjective the scoring of oral reading tests is and the amount of examiner error these scores contain. The score one receives definitely depends upon who is doing the scoring. It is depressingly obvious that scoring comprehension responses if far more difficult than anybody has previously suggested. Both groups--on both protocols--varied between 22 to 43 percentage points in their scoring of the children's answers. This was despite the fact that the answers were written and could be studied and compared with the suggested answers, a situation usually not possible when an administering an oral reading test.

When the study was designed, it was thought that professors of reading might be kind of a criterion group, that their knowledge and experience might serve as a standard with which other less knowledgeable and experienced groups

could be compared. But in scoring comprehension responses, they were just as or even somewhat more variable than teachers. Additional knowledge about and experience with oral reading tests didn't seem to produce less variability, i.e., error.

Determining reading level was also very troublesome for both groups with large numbers of each group choosing both levels--Instructional and Frustration--or both protocols. Professors of reading were much more lenient on both protocols than were teachers. For all practical purposes, 60 percent of the professors rated each protocol as being at the Instructional level while only 21 percent of the teachers did so. I concede that, in borderline cases, legitimate differences of opinion can occur. This however does not decrease the amount of error contained in the scores of oral reading tests--the subject of this study.

Recommendations

The results of oral reading tests are much less precise than we have typically treated them as being. Two sources of error should be considered when interpreting the results of these tests: (1) examinee fluctuation and (2) examiner subjectivity. Either alone is great enough to require us to interpret a score from these tests as occurring not at a point of achievement but within a band of probability. Combining these two sources of error significantly increases the range of probability within which a score is likely to fall.

Thus it seems obvious that we should abandon the concept of a score on an oral reading test as representing a point on a continuum and substitute for

it the more reasonable notion of it being within a band of scores at least a grade level in range. All oral reading tests--standardized or IRIs--should somehow incorporate this concept into the way in which they report their results.

Also, manuals of these tests should make the concept of probability bands much more explicit to test users and should caution against interpreting the results more precisely than is warranted. The fact that the scoring and interpretation of these tests is quite subjective--and how this subjectivity reduces the precision of the scores--deserves far more attention in test manuals than it now receives.

Words such as "about," "approximately," and "estimate" should be widely used when reporting the results of these tests so that people who are less knowledgeable about the magnitude of errors contained in these tests will be less inclined to view them as representing exact points on a continuum of reading growth.

Because answers to comprehension questions in both the easy-to-score and the hard-to-score protocol produced a wide range of opinions as to the percentage of correct responses, authors of oral reading tests should give examiners more definite guidance as to what constitute right and wrong answers--as does the Wechsler Intelligence Scale for Children-Revised.

References

Bradley, Fred, Gerald Hanna, and Bert Lucas. The Reliability of Scoring the WICS-R. Journal of Consulting and Clinical Psychology, 48 (Aug. 1980), 530-531.

Burns, Paul and Betty Roe. Informal Reading Assessment. Chicago: Rand McNally College Publishing Co., 1980.

Durrell, Donald and Jane Catterson. Durrell Analysis of Reading Difficulty, 3rd ed. New York: The Psychological Corp., 1980.

Ekwall, Eldon. Ekwall Reading Inventory. Boston: Allyn and Bacon, Inc., 1979.

Gray, William S. Gray Oral Reading Tests. Indianapolis, IN: Bobbs-Merrill Co., 1963.

Johns, Jerry. Basic Reading Inventory, 2nd ed. Dubuque, IA: Kendall/Hunt Publishing Co., 1981.

Ladd, Eleanor. A Comparison of Two Types of Training with Reference to Developing Skill in Diagnostic Oral Reading Testing. Doctoral dissertation, Florida State University, 1961.

Page, William and Kenneth Carlson. The Process of Observing Oral Reading Scores. Reading Horizons vol. 45 (Spring 1975), pp. 147-150.

Seifert, Mary. Graduate Programs and Faculty in Reading, , 3rd ed. Newark, DE: International Reading Association, 1978.

Silvaroli, Nicholas. Classroom Reading Inventory, 3rd ed. Dubuque, IA: C. Brown Co., 1976.

Spache, George. Diagnostic Reading Scales, Revised Edition, Examiner's Manual, Monterey, CA: CTB/McGraw Hill, 1972, 1981.

Spache, George. Diagnostic Reading Scales, Revised Edition, Technical Bulletin, Monterey, CA: CTB/McGraw Hill, 1975.

Woods, Mary and Alden Moe. Analytical Reading Inventory, 2nd ed. Columbus, OH: Charles E. Merrill Publishing Co., 1981

TABLE 3

Frequency of Percentages in Word Recognition and Comprehension Scores
for Professors of Reading and for Teachers on Easy- and Hard-to-Score Protocols

Protocol													
Easy-to-Score								Hard-to-Score					
Professors				Teachers				Professors				Teachers			
Word Recog.		Comprehension		Word Recog.		Comprehension		Word Recog.		Comprehension		Word Recog.		Comprehension	
Score	Number	Score	Number	Score	Number	Score	Number	Score	Number	Score	Number	Score	Number	Score	Number
91%	5	100%	1	90%	1	86%	5	93%	1	86%	3			93%	1
90%	2	86%	1	89%	2	79%	7	91%	2	79%	3			86%	1
89%	3	79%	3	88%	3	71%	36	90%	3	71%	13	90%	3	79%	11
88%	14	71%	16	86%	67	64%	18	89%	1	64%	4	89%	2	71%	54
		64%	2			57%	7	88%	2	57%	1	88%	69	64%	3
		57%	1					86%	15			86%	1	57%	4
														50%	1
Reading Level								Reading Level					
Instructional . 14				Instructional . 21				Instructional . 15				Instructional . 16	
Frustration . . 8				Frustration . . 53				Frustration . . 8				Frustration . . 57	
Undecided . . . 2				Undecided . . . 0				Undecided . . . 1				Undecided . . . 1	